CORPUS

Corpus

19 | 2019
Corpus et pathologies du langage

Partitionnements multiples de corpus : une lecture polyangulaire ?
L'exemple des bases latines du LASLA

Multiple Corpus: a Polyangular Readings Approach?

Margherita Fantoli et Marc Vandersmissen

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/4505
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Margherita Fantoli et Marc Vandersmissen, « Partitionnements multiples de corpus : une lecture polyangulaire ?
L'exemple des bases latines du LASLA », *Corpus* [En ligne], 19 | 2019, mis en ligne le 01 janvier 2019, consulté le 11 septembre 2019. URL : http://journals.openedition.org/corpus/4505

Partitionnements multiples de corpus : une lecture polyangulaire ? L'exemple des bases latines du LASLA

Multiple Corpus: a Polyangular Readings Approach?

Margherita Fantoli et Marc Vandersmissen

1. Introduction[1]

1 Depuis 1962, le LASLA (Laboratoire d'Analyse Statistique des Langues Anciennes) de l'Université de Liège travaille sur une base de données de textes lemmatisés de langues grecque et latine[2]. Au départ, la lemmatisation avait pour but de faciliter la production d'*Indices verborum*[3] et de concordance de la littérature classique. La plupart d'entre eux ont fait l'objet d'une publication. Ensuite, le LASLA a participé au développement d'outils informatiques et de plateformes statistiques (Hyperbase, Opera Latina, TXM) pour exploiter les données patiemment collectées pour chaque texte lemmatisé. En parallèle, le laboratoire a mis au point une procédure en ligne de lemmatisation semi-automatique validée par un philologue, gage de la qualité scientifique des données ainsi enregistrées[4].

2 Aujourd'hui, la base de textes latins lemmatisés contient environ 2 000 000 de mots dans tous les genres littéraires. Chaque mot d'un texte y est encodé avec son lemme, sa forme dans le texte, sa référence et une analyse morphosyntaxique détaillée. Comparé au corpus latin, les textes grecs traités sont moins nombreux et l'information annotée est moins complète : le lemme, la forme et la catégorie du discours sont attachés à chaque mot du texte. Toutefois, Hyperbase Web Édition a été adapté aux textes de langue grecque[5] et il est désormais possible d'explorer ces bases de manière automatique. De nombreuses études, fondées sur ces outils, ont démontré les possibilités offertes par les méthodes statistiques dans plusieurs domaines : linguistique, littérature, études du discours[6]

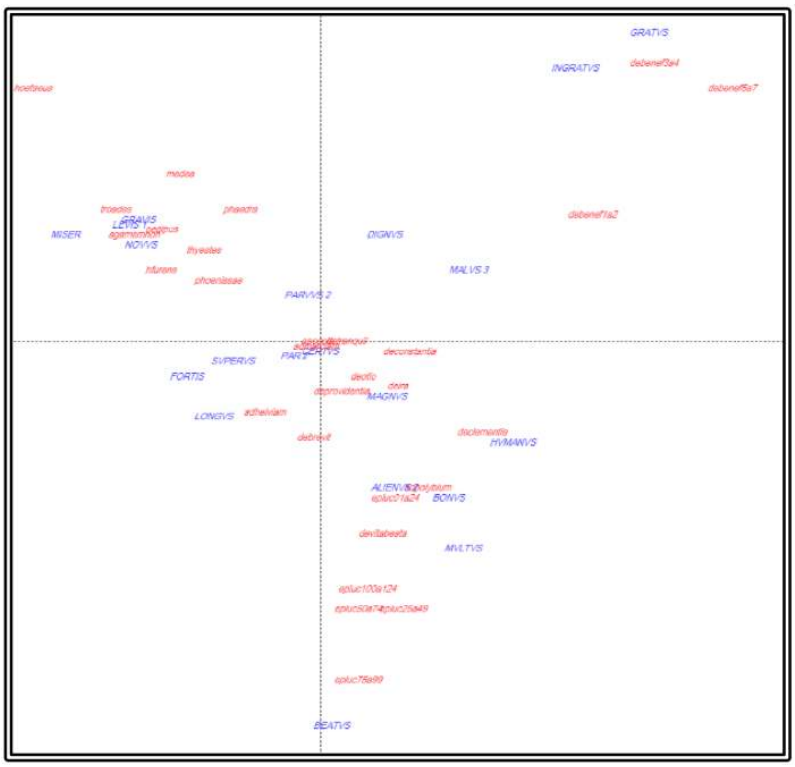
3 Rapidement, le LASLA a inscrit ses travaux dans le contexte théorique de l'Analyse de Données Textuelles (ADT). Ses chercheurs partagent avec ce domaine un intérêt méthodologique sur les conséquences de l'approche statistique sur la définition du texte lui-même. Sur la base d'une réflexion sur cette approche herméneutique, les concepts de lexicométrie, textométrie et enfin logométrie ont été définis avec la création de nouveaux outils statistiques[7]. Ces méthodes prennent en compte toujours plus d'aspects de la réalité textuelle, des comptages lexicaux aux mesures du discours. Dans ce cadre, nous aimerions proposer un concept complémentaire afin de nourrir la recherche : la lecture polyangulaire.

2. Ordinateur et textes : lecture linéaire et lecture réticulaire

4 En 2005, Jean-Marie Viprey montrait que la lecture humaine était linéaire alors que la lecture numérique était tabulaire ou réticulaire[8]. En effet, le format du texte influence sa compréhension. Puisque l'être humain lit le texte ligne après ligne, il l'envisage d'abord comme une séquence continue de mots alignés. À l'opposé, l'ordinateur permet de s'affranchir des contraintes de la linéarité en comptant, à travers le texte, différents éléments de manière verticale (mots, groupes de mots, motifs lexicaux ou syntaxiques, objets grammaticaux...). L'ordinateur repère automatiquement les données étudiées par le chercheur et ouvre des fenêtres localisées dans le texte, par exemple dans le cas de concordances. Cette approche induit une nouvelle perception de la matérialité du texte en présentant les résultats sous la forme de tableaux ou de matrices.

5 Le développement des matrices marque un moment important dans le domaine de la statistique de corpus. La matrice ordonne l'information à traiter en lignes représentant les mots cherchés et en colonnes classant les textes présents dans le corpus. L'élément de la matrice (*i, j*) désigne donc la fréquence absolue ou relative – ou encore éventuellement la spécificité – d'un mot figuré par la ligne *i* dans le texte représenté par la colonne *j*. À partir de ces données ainsi hiérarchisées, il est possible d'effectuer différents types de calcul (AFC, Analyse arborée), qui visent généralement à représenter la proximité ou la distance des textes du corpus à partir de la ressemblance plus ou moins forte des colonnes correspondantes. On peut également étudier l'affinité entre les mots recherchés dans le corpus et la similarité entre les lignes. Par exemple, l'AFC de la distribution des vingt adjectifs les plus fréquents chez Sénèque montre clairement le regroupement des textes dramatiques autour de *miser, grauis, nouus, leuis*, ainsi que la proximité des *Lettres à Lucilius*, caractérisées par leur emploi de l'adjectif *beatus*. Notons aussi que les livres du *De Beneficiis* se rassemblent autour du couple *gratus* vs *ingratus*.

Figure 1. AFC des 20 Adjectifs les plus fréquents de la base SENECA



6 Relevons que ces démarches sont purement 'quantitatives'. Ce qui compte est la présence (et le poids de cette présence) du mot dans le texte. Le contexte et tous les éléments propres à une lecture linéaire ou même réticulaire (qui considère les passages dans lesquels les mots cherchés se trouvent) ne sont pas pris en compte[9]. Il s'agit donc d'une approche radicalement différente des précédentes qui demande un outillage interprétatif nouveau et qui puisse concilier ce type d'information avec le retour au texte.

7 Ces évolutions n'auraient pas été possibles sans le passage du texte manuscrit au texte numérique et sans le développement de nouveaux outils d'interrogation. L'ordinateur permet aussi d'étudier des corpus toujours plus importants au risque de rendre impossible toute lecture linéaire du texte et de déconnecter l'interprétation des données de la réalité textuelle. Pour éviter cette éventualité, durant les JADT de 2006, J.-M. Adam a suggéré de concilier les deux approches pour une compréhension globale des textes comme « la combinaison de parcours linéaires et réticulaires »[10]. L'ordinateur facilite cette approche car il permet non seulement d'extraire des résultats détaillés mais aussi d'offrir un lien permanent vers le contexte. Il permet de passer des résultats statistiques aux textes eux-mêmes. Il est ainsi possible pour le chercheur de fonder ses conclusions sur une connaissance du texte, même dans de larges corpus, les approches quantitatives et qualitatives étant par là réunies.

3. Ordinateurs et corpus : lecture polyangulaire ?

8 De plus, l'utilisation de l'ordinateur n'a pas seulement influencé nos pratiques de lecture, il a aussi fait évoluer notre conception du corpus de texte. Comme D. Mayaffre l'a étudié, constituer un corpus digital nécessite une démarche plus contraignante. Il n'est plus

question d'assembler des textes qui feront plus ou moins partie de la recherche en fonction des résultats. Les textes font désormais l'objet d'un processus méticuleux de sélection et seront intégrés ou non au corpus étudié ensuite[11]. Avec sa dématérialisation, le corpus devient paradoxalement un espace fermé. C'est dans cet ensemble délimité que l'ordinateur peut effectuer ses opérations statistiques. Mais même après cette évolution, les chercheurs, qui travaillent sur ces corpus préalablement définis, ont recours aux processus d'analyse linéaire (philologie, linguistique) et/ou réticulaire (linguistique de corpus, ADT) ou en associant les deux approches, comme indiqués ci-avant.

9 Aujourd'hui, les logiciels – nous prendrons l'exemple d'HYPERBASE WEB ÉDITION avec lequel nous travaillons – permettent non seulement d'appliquer des recherches documentaires (index, concordances) ou statistiques (spécificités, AFC) avec un retour au contexte. Mais il offre aussi – et c'est assez nouveau – des outils de construction de corpus. Lorsque vous avez sélectionné le corpus dans lequel vous souhaitez effectuer une recherche, il vous est désormais possible de l'articuler de plusieurs manières sans procédure informatique lourde, comme par le passé. Dans le corpus sélectionné, on peut conserver la division en œuvres proposée par le logiciel. Mais il est surtout possible de constituer facilement des ensembles différents en fonction des besoins du chercheur. Cette fonctionnalité repose sur l'emploi des métadonnées dans le corpus. Chaque mot y est enregistré avec son analyse morphosyntaxique, ce sont les données. Mais il peut également être attaché à un certain nombre d'informations complémentaires au sujet de ces données, ce sont les métadonnées : auteur, œuvre, chapitre, période... L'enrichissement du corpus avec des métadonnées permet de sélectionner le ou les textes en fonction du critère souhaité. Le logiciel propose ainsi plusieurs possibilités :

-exclure une partition : si un texte s'écarte trop de l'ensemble, il peut être neutralisé pour un corpus plus homogène. Par exemple, si on travaillait sur un corpus dont on doute de l'authenticité de certaines œuvres, on peut facilement ajouter ou exclure ces œuvres de l'ensemble étudié.

-fusionner des partitions : il est possible d'unir des partitions pour faire apparaître d'autres critères de recherche. Par exemple, on peut réunir des œuvres qui traitent d'un même cycle mythologique ou par période chronologique.

-nouvelle division : si les métadonnées ont été préalablement encodées, on peut également diviser différemment le corpus, selon un critère différent. Par exemple : par auteur, par œuvre, par genre.

10 Ainsi, tout en travaillant sur un même ensemble de textes constituant un corpus fermé, son partitionnement peut s'opérer selon plusieurs critères : genre littéraire, période chronologique, auteur... Grâce au jeu des métadonnées, le chercheur peut envisager plusieurs partitionnements/ constructions de corpus destinés à faire apparaître des réalités linguistiques différentes. C'est en effet par la comparaison entre les sous-corpus que les particularités textuelles de chacun d'entre eux sont mises en évidence, c'est l'exploration de corpus contrastive. Dans la mesure où certaines précautions méthodologiques sont prises préalablement (sous-corpus équilibrés, homogénéité du corpus), la validité de cette approche n'est plus à démontrer[12]. Mais, l'ordinateur permet maintenant de faire varier à souhait, au sein d'un même corpus, les sous-corpus qui le composent.

11 Avec ce développement, il devient donc nécessaire de multiplier les constructions de corpus pour dégager le plus d'informations possible sur ce corpus et pour en appréhender sa complexité textuelle et linguistique. Cette nouvelle possibilité, d'abord technique,

modifie à nouveau notre relation au corpus, et surtout doit nous faire prendre conscience qu'un changement épistémologique est nécessaire. Le corpus n'est plus un terrain d'exploration figé qui se traverse seulement de manière horizontale (lecture linéaire) ou verticale (réticulaire). À la place, il peut être abordé sous plusieurs angles, selon plusieurs points de vue qui éclaireront chacun des phénomènes linguistiques ou textuels différents. Le corpus peut être approché, non plus seulement, comme l'objet où se déroule la recherche, mais aussi comme un outil permettant de l'explorer lui-même. L'évolution technique fait donc changer le corpus de statut ontologique. Il passe d'un objet de recherches à sujet de recherches.

12 Les résultats obtenus des différentes configurations du corpus sont ensuite compulsés et mis en relation pour enrichir notre compréhension globale du texte. C'est ce que nous appellerons la lecture polyangulaire. Nous aimerions ainsi que ce concept puisse s'ajouter à la boite à outils théorique à disposition du chercheur et venir compléter les principes de lectures linéaire, réticulaire ou matricielle. De plus, la lecture polyangulaire dépasse le concept de *corpus driven* car il ne s'agit pas seulement de se laisser guider par le corpus. Elle vise à parcourir volontairement le plus possible le corpus sélectionné au départ pour obtenir le plus de résultat possible. Dans cette démarche, le chercheur occupe une place active dans le processus de la lecture polyangulaire. L'interrogation multiple devient une fin en soi. En d'autres termes, si le texte est envisagé comme un maillage multidimensionnel, la lecture polyangulaire vise à mettre au jour le plus de mailles possible afin d'améliorer notre connaissance du tissu textuel étudié. La maille se joue ainsi au niveau de la séquence linaire du texte, mais peut également intervenir au niveau de l'intertexte et prendre son sens au-delà de la simple cooccurrence directe[13].

4. Lecture polyangulaire : deux exemples d'application

13 Pour dépasser le raisonnement strictement théorique, proposons deux exemples : l'un sur la tragédie de Sénèque et l'autre sur les historiens latins. Le premier s'intéresse au théâtre de Sénèque et plus particulièrement au sentiment de *furor*, « la folie furieuse », qui structure ces œuvres dramatiques[14]. Il s'agit en effet d'un des principaux états émotionnels qui mènent le héros à commettre le crime tragique au centre de chaque pièce : l'infanticide pour Médée, l'anthropophagie pour Thyeste, le meurtre de l'époux pour Clytemnestre... Nous travaillons donc dans un corpus fermé constitué de neuf tragédies[15]. Grâce aux outils dont nous disposons aujourd'hui, on peut non seulement chercher ce terme clef dans le corpus. Mais surtout, il est possible d'en étudier sa distribution selon plusieurs divisions de corpus en appliquant le principe de la lecture polyangulaire. Ainsi, nous proposons de lire les trois histogrammes suivants illustrant la distribution du lemme *furor* dans la tragédie de Sénèque selon trois critères différents : par tragédie, par catégorie de personnages et par personnage.

Figure 2. Histogramme de la distribution du lemme *furor* entre les neuf tragédies de Sénèque

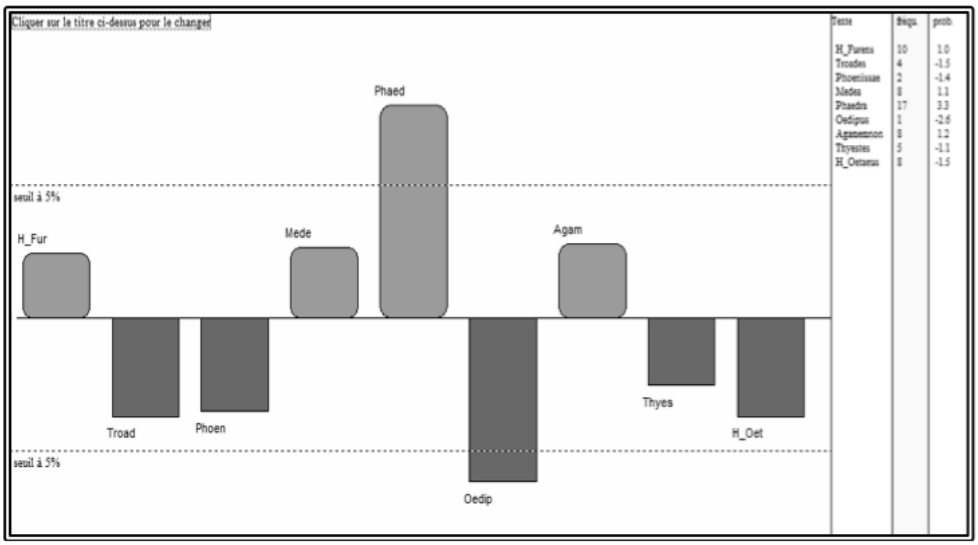


Figure 3. Histogramme de la distribution du lemme furor entre cinq catégories de personnages tragiques (chœurs, messagers, nourrices, héros masculins et héros féminins)

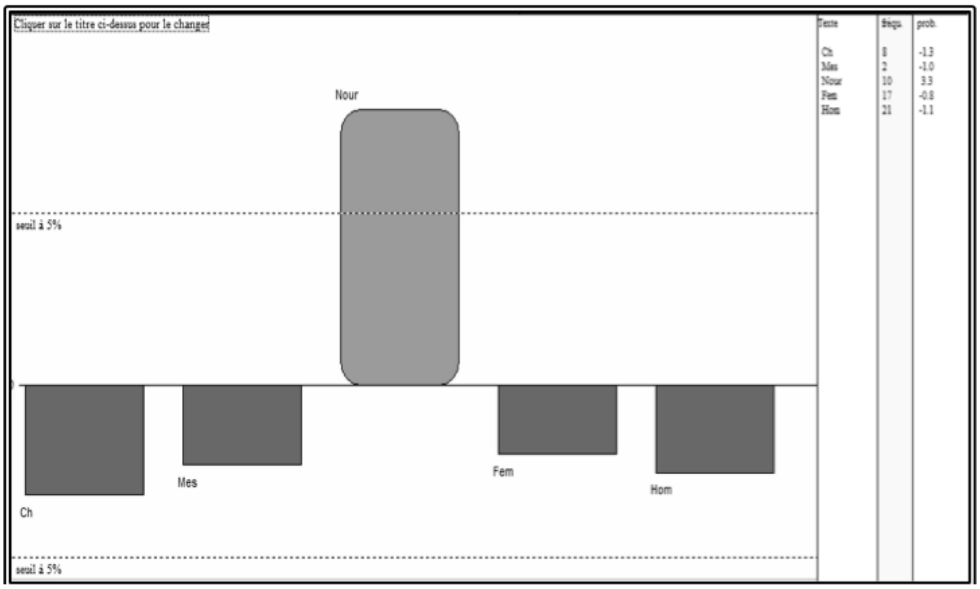
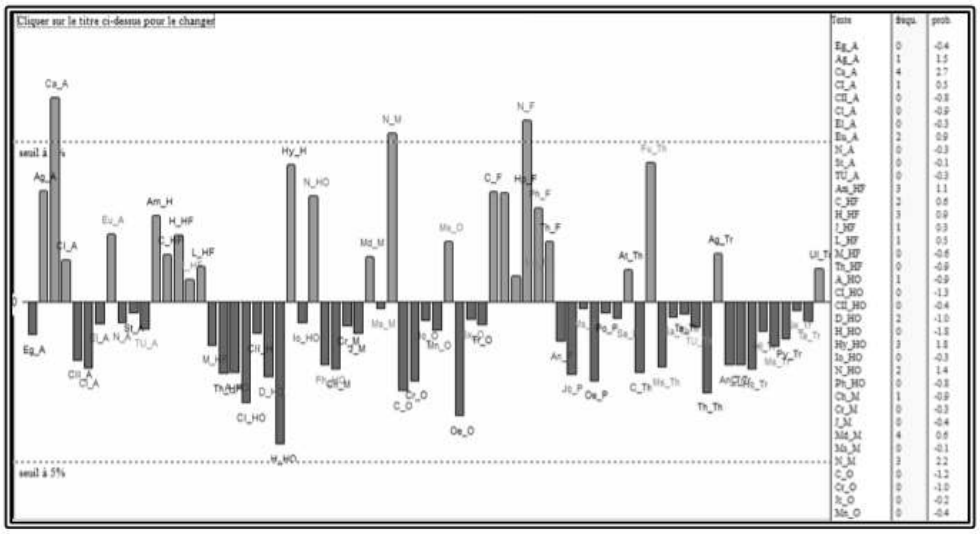


Figure 4. Histogramme de la distribution du lemme *furor* entre les 71 personnages tragiques



14 Alors que le terme apparaît dans toutes les tragédies, le premier histogramme nous apprend que *furor* est surutilisé dans *Phèdre*. Cette particularité pourrait s'expliquer par la trame même de cette pièce. À la différence des autres tragédies, les trois personnages principaux de celle-ci, à savoir Hippolyte, Phèdre et Thésée, vivent le développement tragique du *dolor – furor – nefas*. Phèdre est le membre pivot de l'action dramatique[16] mais elle provoque également le processus chez son beau-fils et ensuite chez son époux (*Phèdre* , v. 1164-1169). Le terme est donc beaucoup utilisé par les héros, plus nombreux à être concernés par le *furor* que dans les autres pièces. Cette observation est également renforcée par le troisième histogramme qui montre que tous les personnages dépassent la moyenne d'utilisation de ce terme. À l'inverse, l'*Œdipe* présente une sous-utilisation significative du lemme. Ce résultat confirme la lecture de F. Dupont[17]. Pour la philologue, le *nefas* d'Œdipe a été commis préalablement à la pièce lorsque le héros a tué son père et a épousé sa mère. La tragédie se concentre sur la découverte du crime et sa révélation sur scène. Dans ce cas, le *furor* est donc indirect et n'est pas au cœur du développement des héros, ce qui explique le faible nombre d'occurrences du terme dans cette pièce.

15 Le deuxième histogramme apporte un angle de vue différent mais complémentaire au premier. De manière moins attendue, il indique que la catégorie de personnage « nourrices » surutilise largement le terme *furor*. En effet, les nourrices ne sont jamais elles-mêmes sujettes au *furor* à l'inverse des personnages masculins et féminins. Dans le théâtre gréco-romain, elles constituent des personnages secondaires à l'action tragique proprement dite. Elles sont les corps d'une identité générique et ne sont pas identifiées par un nom qui leur serait propre. Or elles suremploient significativement un terme attaché aux héros tragiques. Le troisième histogramme nous indique même quelles nourrices participent le plus à ce résultat : les nourrices de Médée (*Médée*) et de Phèdre (*Phèdre*) ainsi que, dans une moindre mesure, celle de Déjanire (*Hercule sur l'Œta*). Seule la quatrième nourrice du corpus (*Agamemnon*) ne renforce pas cette tendance. Le rôle à la fois de miroir et de catalyseur de la nourrice devant les héroïnes a déjà été démontré[18]. Elles aident leur protégée à mettre des mots sur leur développement émotionnel et à faire face à leur passion ; elles contribuent par là à les faire passer du *dolor* vers le *nefas* (voir par exemple Médée-nourrice : *Médée*, v. 150-178). L'histogramme montre ici que le *furor* imprègne le discours des nourrices parce qu'elles sont les premiers témoins directs de ce

sentiment qui naît chez les héroïnes. Elles le décrivent donc avec beaucoup d'inquiétude mais aussi beaucoup de précision.

16 Comme dit ci-avant, le troisième histogramme illustre l'utilisation significative de *furor* par les nourrices. Il indique aussi un suremploi du terme chez Cassandre dans l'*Agamemnon*. Il est intéressant de relever cette particularité. En effet, Cassandre, ancienne princesse troyenne devenue l'esclave du roi grec Agamemnon après la prise de Troie, n'est pas le personnage central de la pièce. Ici, c'est le couple adultère Clytemnestre-Égisthe qui suit le développement tragique traditionnel jusqu'au meurtre d'Agamemnon. Néanmoins, en raison de sa qualité de devineresse, Cassandre est prise de *furor* lorsque le dieu Apollon lui envoie une vision ou une prophétie (*Agam.*, v. 720-725). C'est dans cet état émotionnel qu'elle est capable de recevoir et d'interpréter les messages divins. Contrairement aux autres mortels, la princesse troyenne est à mi-chemin entre le monde des hommes et celui des dieux. Son intervention dans l'*Agamemnon* consiste principalement en un message de type oraculaire. Cette caractéristique marque donc aussi son discours, comme démontré ici avec la spécificité du lemme *furor*.

17 Cet exemple sur la tragédie latine démontre comment la lecture polyangulaire permet de faire émerger des résultats différents mais complémentaires. Ils enrichissent notre compréhension du concept de *furor*. Au-delà d'un aspect thématique, le partitionnement de corpus en catégorie de personnages et en personnages permet d'atteindre la construction dramatique même des pièces et la caractérisation de certains personnages. Il y aurait donc lieu d'appliquer encore cette approche pour poursuivre notre exploration du corpus sénéquéen.

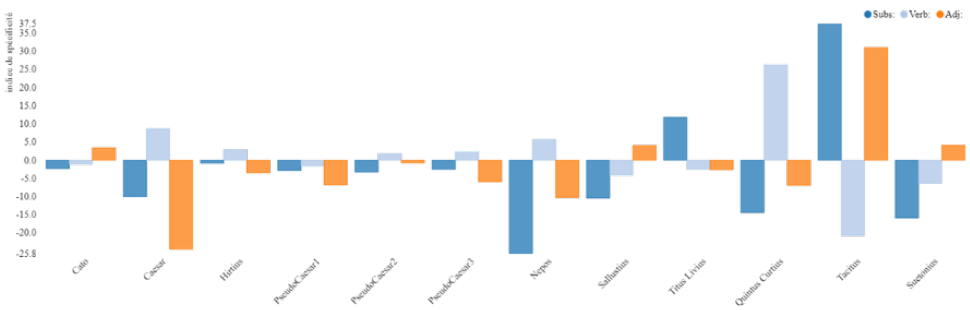
18 Le deuxième exemple a pour but de montrer comment l'analyse d'aspects syntaxiques d'un corpus d'auteurs peut donner des résultats différents grâce à l'emploi des fonctionnalités de partitionnement de corpus mises à disposition par Hyperbase Web Édition. La base HISTORIA, proposée par le site, contient les œuvres principales de l'historiographie latine, organisées en ordre chronologique pour faciliter l'interprétation des données pour l'utilisateur. La partition proposée par Hyperbase Web Édition contient les œuvres suivantes, divisées par livres : les *Origines* de Caton, les œuvres de César (*Bellum Gallicum* et *Bellum Ciuile*), *Bellum Africum, Bellum Alexandrinum, Bellum Hispaniense,* les œuvres de Salluste (*De coniuratione Catilinae*, *Bellum Iugurthinum, Fragmenta*), les *Vitae* de Cornelius Nepos (réunies en une seule partition pour des raisons de taille), la première décade de Tite-Live, les livres de 3 à 10 des *Historiae Alexandri Magni* de Quinte-Curce, l'*Agricola* de Tacite, les livres 1-5 des *Historiae,* les livres 1-6 et 11-16 des *Annales*, les *Vitae Caesarum* de Suétone, où chaque vie représente une partition.

19 La fonction « édition » permet de se déplacer facilement à travers trois niveaux : les auteurs, les œuvres et les livres. En effet, en plus du partitionnement proposé automatiquement par la base, il est aisé de regrouper toutes les œuvres d'un auteur en une seule partition, ou encore tous les livres d'une œuvre en une seule partition. L'analyse de la distribution de certains phénomènes linguistiques peut donc être faite sur les trois niveaux et nous apporte ainsi trois informations différentes. En premier lieu, la comparaison entre les auteurs fait apparaître deux aspects : d'un côté la présence ou non de spécificités d'auteur à l'intérieur d'un même genre littéraire (l'historiographie dans ce cas) ; de l'autre côté l'existence ou non d'une évolution chronologique des traits linguistiques sur lesquels porte l'interrogation. En deuxième lieu, la comparaison entre les œuvres permet d'attirer l'attention sur la distance entre les œuvres d'un même auteur, et peut également indiquer si certaines œuvres d'auteurs différents montrent des

affinités (ceci pourrait par exemple fournir des données intéressantes sur le rapport entre œuvres d'imitation et modèles). En troisième lieu, la démarche de confronter ensemble chaque livre répond à la question de l'homogénéité interne des œuvres (par exemple, pour des œuvres dont la production s'étale sur une longue période, ceci pourrait fournir des indications sur l'évolution du style de l'auteur). Ce processus pourrait également mettre en évidence que certaines parties des travaux montrent des spécificités de style (parties introductives ou finales etc.).

20 Les possibilités d'interrogation pour analyser l'aspect syntaxique de la langue sont infinies. On préfère ici se concentrer sur l'ADN de constitution de la phrase, en considérant la distribution des parties des discours. L'équilibre entre l'emploi de verbes, substantifs, adverbes, pronoms fournit une première information sur le style du texte analysé[19]. Le premier histogramme, à partir de la catégorie « auteur » (obtenu en fusionnant les livres du même auteur dans la partition initiale proposée par Hyperbase Web Édition), donne le résultat suivant :

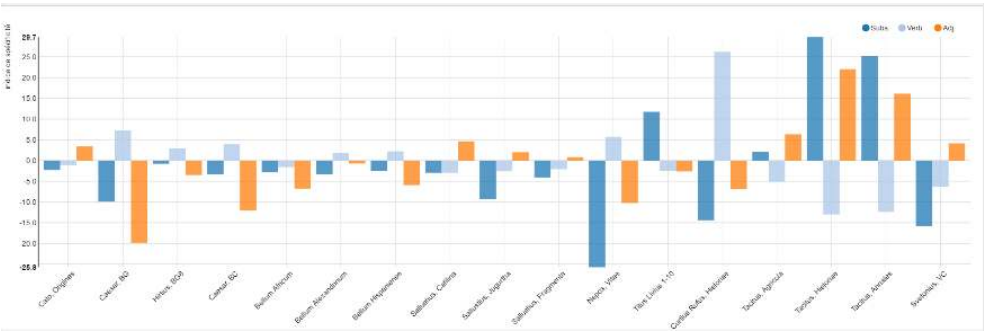
Figure 5. Histogramme de la distribution de verbes, substantifs, adjectifs dans les auteurs de la base HISTORIA



21 Signalons d'abord que la dimension de certaines partitions pourrait écraser les résultats des auteurs dont l'étendue des textes est fortement inférieure à celle des autres. Toutefois, il est évident que les auteurs les plus représentés dans la base montrent des spécificités assez hautes, dont la valeur absolue dépasse 10 (les résultats doivent être ici considérés comme significatifs à partir de 5). Ceci signifie que la balance entre les parties du discours varie fortement d'un auteur à l'autre : si Tacite montre une forte prédilection pour substantifs et adjectifs et un sous-emploi des verbes, l'inverse est constaté pour Quinte-Curce. Par ailleurs, César montre un style particulièrement pauvre en adjectifs ; chez Tite-Live on voit une affinité pour les substantifs, au contraire de Cornelius Nepos. Ainsi, même dans le cadre d'un seul genre littéraire, les données montrent que le style peut varier fortement. De plus, on voit aisément que le facteur chronologique n'est pas décisif, car aucune évolution régulière ne ressort de l'histogramme. Remarquons néanmoins que les derniers auteurs – du moins Quinte-Curce et Tacite – présentent des très hautes valeurs positives : le premier pour les verbes et le second pour les substantifs et les adjectifs. Même si les spécificités positives portent sur des catégories du discours différentes, on pourrait se demander si les auteurs de cette période ont une tendance à déséquilibrer leur prose par le suremploi d'une certaine catégorie.

22 Passons maintenant à l'analyse du deuxième histogramme. Le corpus est désormais partitionné par œuvre, ce qui s'obtient en fusionnant les livres d'une même œuvre à partir de la division initiale proposée par Hyperbase Web Édition :

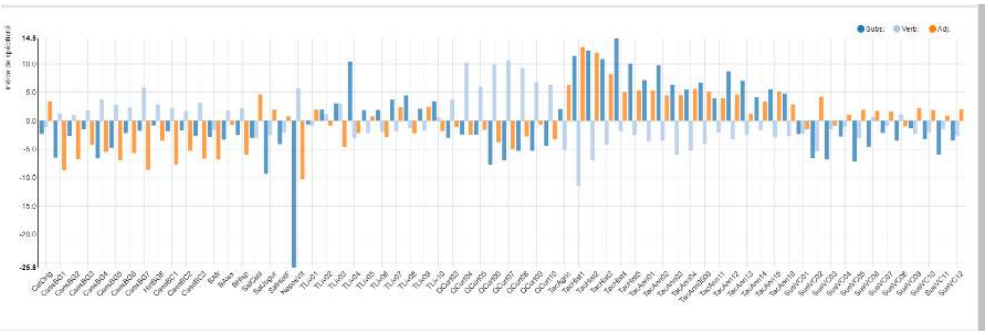
Figure 6. Histogramme de la distribution de verbes, substantifs, adjectifs dans les œuvres de la base HISTORIA



23 On peut remarquer certains traits intéressants. Les œuvres d'un même auteur tendent à avoir des proportions similaires entre les parties du discours. Ceci se vérifie chez les trois auteurs dont la base contient plusieurs œuvres : César, Tacite et Salluste. La précision des résultats est frappante. Ceci induit que l'équilibre entre les parties du discours est donc une partie essentielle de la façon de construire la phrase des auteurs analysés. Il est aussi intéressant de remarquer que le huitième livre du *Bellum Gallicum* montre des proportions très proches de celles des œuvres de César, compte tenu de la taille de cette partition de faible étendue, comparée au reste du *Bellum Gallicum*. Ceci pourrait soutenir l'hypothèse que cet aspect de la langue de l'auteur devient, aux yeux des continuateurs de l'œuvre, une partie intégrante du style à imiter[20].

24 Dans le troisième histogramme, l'analyse par livre confirme la même impression :

Figure 7. Histogramme de la distribution de verbes, substantifs, adjectifs dans les livres des œuvres de la base HISTORIA



25 D'abord, on remarque que les livres de Quinte-Curce et Tacite ont la même tendance aux hautes spécificités montrée jusqu'à présent : il s'agit donc bien d'une donnée réelle et pas de l'effet d'un déséquilibre dans la dimension des partitions. Ensuite, les œuvres montrent une forte homogénéité interne : on remarque un certain degré de variation chez Tite-Live (et de façon mineure en Suétone), mais il s'agit principalement de valeurs trop basses pour être statistiquement significatives. Au vu de l'équilibre entre les partitions, on remarque que la distribution des adjectifs, verbes et substantifs dans le huitième livre de la *Guerre des Gaules* est totalement cohérente avec celle des livres précédents. Le *Bellum Hispaniense* montre la même similitude avec les œuvres de César.

26 Cette approche nous permet ainsi de constater de façon assez rapide quels sont les critères les plus intéressants pour l'étude de l'aspect qu'on a isolé. Le critère chronologique ne semble pas donner beaucoup de résultats. Toutefois on pourrait se

demander si, en poussant les limites chronologiques plus en avant, la tendance au suremploi de certaines parties du discours persiste ou non. Il ressort clairement que le facteur *auteur* est le plus puissant, car on remarque une forte cohérence entre les œuvres d'un même auteur, ainsi qu'à l'intérieur d'une même œuvre. Une étape complémentaire dans l'étude pourrait être (en choisissant une base plus ample) la question du genre littéraire. Que se passe-t-il en mettant dans une même base des œuvres de genres littéraires différents ? Les œuvres du même genre montrent-elles une affinité particulière, ou la différence entre les auteurs reste-t-elle prédominante ? Que se passe-t-il avec des auteurs qui écrivent des œuvres de genres littéraires différents ? De plus, la cohérence dans la distribution des parties du discours invite à penser qu'il s'agit d'un facteur important à prendre en compte dans l'étude de la constitution de la phrase d'un auteur. La question de savoir comment ceci se reflète dans les effets stylistiques pourrait être exploitée en précisant la recherche.

27 En résumé, cet exemple montre de façon claire que la lecture polyangulaire repose à la fois sur le *corpus driven* et le *corpus based* : l'interrogation était prédéterminée, ainsi que l'ensemble du corpus. Toutefois, le critère le plus intéressant pour partitionner le corpus, et donc la perspective dans laquelle considérer la question, a été déterminé sur la base des résultats. Ainsi les étapes successives de l'étude ont été suggérées par ces mêmes résultats.

5. Conclusions

28 En guise de conclusions, nous espérons avoir réussi à démontrer l'intérêt d'appliquer la lecture polyangulaire lors de recherches linguistiques, textuelles ou discursives. Cette démarche invite le chercheur à explorer son corpus selon plusieurs angles, à savoir plusieurs partitionnements, afin d'en découvrir le plus d'aspects possibles. C'est par la multiplication et le croisement des partitionnements que pourront émerger les facteurs les plus caractérisants des textes étudiés. Située entre les méthodes de *corpus based* et *corpus driven*, la lecture polyangulaire vient compléter les notions de lectures linéaire, réticulaire et matricielle d'un corpus. Ensemble, ces outils permettent de mettre au jour des réalités et des mécanismes encore inconnus à tous les niveaux de la textualité. De plus, cette démarche, inscrite dans un retour permanent au texte, tend à réduire la part de subjectivité inhérente à un choix de critères opérés *a priori*. Le corpus devient lui-même un outil d'interrogation de corpus ; il est à la fois objet et sujet d'étude pour mieux en appréhender sa complexité. L'amélioration des outils informatiques permet ainsi de faciliter cette procédure et d'enrichir toujours plus notre connaissance des textes. Une fois encore, la création de nouveaux supports questionne notre propre méthode de travail et encourage à repenser notre rapport au texte.

BIBLIOGRAPHIE

Adam J.-M. (2006). « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles », conférence d'ouverture aux JADT 2006, p. 5 [texte en ligne sur Lexicométrica (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)].

Buffa M. (1986). « Struttura e stile di B.G. VIII », *SRIC* 7 : 19-49.

Canali L. (1966). « Osservazioni sul Corpus cesariano », *Maia* 18 : 115-137.

Delatte L., Évrard Ét., Govaerts S., Hazette P. (éd.) (1962). Sénèque, Consolation à Polybe. Index verborum, relevés statistiques. La Haye : Mouton.

Delatte L., Govaerts S., Denooz J. (1978). L'ordinateur et le latin. Techniques et méthodes, morphologie, syntaxe, lexicologie. Stylistique. Liège : LASLA.

Denizot C., Vandersmissen M. (2018). « De nouvelles perspectives pour les textes grecs du LASLA : Avancées dans la banque de données », *Euphrosyne* 46 (sous presse).

Dupont F. (1995). *Les monstres de Sénèque : pour une dramaturgie de la tragédie romaine*. Paris : Belin.

Gaertner J. F. (2018). « The *Corpus Caesarianum* », *in* Grillo L., Krebs C. B. (éd.) *The Cambridge Companion to the Writings of Julius Caesar*. Cambridge, 263-276.

Longrée D., Mellet S. (2009). « Syntactical Motifs and Textual Structures : Considerations based on the Study of a Latin Historical Corpus », *Belgian Journal of Linguistics* 23 : 161-173.

Longrée D., Mellet S. (2013). « Le motif, une unité phraséologique englobante : Étendre le champ de la phraséologie de la langue au discours », *Langage* 189 : 65-70.

Malingrey A.-M. (éd.) (1978). Indices Chrysostomici I, Ad Olympiadem, Ab exilio epistula, De providentia Dei. Hilsdesheim : Olms.

Marcucci S. (1997). *Analisi e interpretazione dell'*Hercules Oetaeus. Rome : Istituti Editoriali e Poligrafici Internazionali.

Mayaffre D. (2005). « De la lexicométrie à la logométrie », *Astrolabe,* 1-11.

Mayaffre D. (2007). « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? », *in* Rastier F., Ballabriga M. (dir.) *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVII^e Colloque d'Albi Langages et Signification*. Toulouse : Presses Universitaires de Toulouse, 15-25.

Mayaffre D. (2009). « L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle (partie I) », *Lexicometrica Topographie et topologie textuelles*, p. 2-12.

Mellet S., Rollinat-Levasseur E.-M. (1989). « Sénèque, Phèdre : Remarques à propos de la personne », *Vita Latina* 116 : 37-42.

Mellet S. (1998). « Les tragédies de Sénèque vues à travers Hyperbase », *in* Mellet S., Vuillaume M. (éd.), *Mots chiffrés et déchiffrés : Mélanges offerts à Étienne Brunet*. Paris, 255-271.

Poudat C., Landragin F. (2017). *Explorer un corpus textuel. Méthodes. Pratiques. Outils.* Louvain-la-Neuve.

Purnelle G. (2012). « Vers long, vers court, lieux de différenciation lexicale et linguistique : le cas du distique élégiaque latin », *in* Dister A., Longrée D., Purnelle G. (éd.), *JADT 2012 – Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles*. Liège : LASLA : 805-819.

Richter W. (1977). *Caesar als Darsteller seiner Taten : Eine Einführung.* Heidelberg.

Vandersmissen M. (2019). *Le discours des personnages féminins chez Sénèque : approches logométriques et contrastives d'un corpus théâtral*. Bruxelles : Latomus (sous presse).

Viprey J.-M. (2005). « Philologie numérique et herméneutique intégrative », *in* Adam J.-M., Heidmann U. (éd.), *Sciences du texte et analyse de discours*. Genève : Slatkine, 51-68.

NOTES

1. Cet article est une version adaptée de la conférence « Multiple Corpus : a Polyangular Readings Approach? » présentée à la University of Latvia lors du colloque *A corpus and usage-based approach to Ancient Greek* (Riga, 14/04/2018).
2. http://web.philo.ulg.ac.be/lasla/
3. Delatte L., Évrard Ét., Govaerts S. et Hazette P. (éd.) 1962 ; Malingrey A.-M. (éd.) 1978.
4. Delatte L., Govaerts S., Denooz J. 1978.
5. Denizot C., Vandersmissen M. 2018.
6. Mellet S. 1998 ; Longrée D., Mellet S. 2009 ; Purnelle G. 2012 ; Longrée D. – Mellet S. 2013.
7. Mayaffre D. 2005.
8. Viprey J.-M. 2005.
9. Le calcul des cooccurrences permet dans une certaine mesure de mettre les deux approches en relation. Cf. Poudat C., Landragin F. 2017 : 200-209.
10. Adam J.-M. 2006 : 5.
11. Mayaffre D. 2007 : 20-22.
12. Mayaffre D. 2009 : 3-4.
13. Adam J.-M. 2006 : 7-13.
14. Sur le développement dramatique *dolor*, *furor*, *nefas*, voir par exemple : Dupont F. 1995.
15. *Phèdre*, *Médée*, *Œdipe*, les *Phéniciennes*, les *Troyennes*, *Agamemnon*, *Thyeste*, *Hercule Furieux* et *Hercule sur l'Œta*. Nous avons conservé cette dernière même si son authenticité n'est pas strictement assurée. En effet, de cette manière, on travaille sur un corpus homogène et exhaustif puisqu'il réunit l'ensemble des tragédies latines à thème mythologique conservées.
16. Mellet S., Rollinat-Levasseur E.-M. 1989.
17. Dupont F. 1995 : 211-212.
18. Marcucci S. 1997: 241. Vandersmissen M. 2018.
19. Une analyse complète demanderait de prendre en compte les huit parties du discours traditionnellement identifiées. Ici, le but étant seulement d'illustrer l'utilité de la lecture polyangulaire, nous nous limiterons aux catégories suivantes : verbes, substantifs et adjectifs.
20. Dans la préface au livre VIII Hirtius s'excuse de ne pas avoir l'élégance de style 'naturelle' propre aux œuvres de César, tout en annonçant vouloir en continuer le travail. Le style du huitième livre a été comparé à celui du reste de l'œuvre par des nombreuses études. Pour une synopsis récente de la question, cf. Gaertner J. F. 2018. En particulier Buffa M. 1986 ; Canali L. 1966 ; Richter W. 1977. Ce dernier, à propos de l'effort d'Hirtius de se conformer au style de César, écrit (p. 197) : «*Daß hier ein anderer Mann als Caesar schreibt, is bei aller Anlehnung an diesen schon an seiner Sprache zu erkennen. Er gebraucht Wörter und Ausdrücke, die Caesar in allen seinen Büchern gemieden*

hat. Er kennt gelegentlich ungewöhnliche Konstruktionen und bedient sich militärischer Fachausdrücke meist unbefangener als dieser. Aber bei alledem ist er doch ziemlich zurückhaltend; man darf annehmen, daß er sich große Mühe gab, sein Buch so unauffällig wie möglich und unter bewußtem Verzicht auf persönliche Stilentfaltung zwischen Caesars Bücher einzufügen».

RÉSUMÉS

Dans le domaine de l'analyse des données textuelles (ADT), les chercheurs s'intéressent à la relation entre le texte et son support d'exploration. Ces dernières années, l'évolution de l'informatique a profondément modifié notre rapport au texte induisant le développement de nouveaux outils d'étude et critères d'analyse. Dans ce contexte théorique, le concept de lecture polyangulaire permet de compléter les notions de lectures linéaire, réticulaire ou matricielle. Cette approche du texte est devenue possible grâce aux outils d'édition de corpus toujours plus performants proposés par les logiciels de traitement automatique des textes, tel Hyperbase Web Édition. Les fondements de la lecture polyangulaire reposent à fois sur les techniques de *corpus based* et de *corpus driven*. Elle vise à multiplier les partitionnements de corpus afin de faire émerger aussi précisément que possible la trame du tissu textuel. Cette nouvelle méthode est ensuite illustrée par deux exemples issus des bases latines diffusées par le LASLA de l'Université de Liège.

In the field of the Analysis of Textual Data (ATD), the researchers are interested in the relation between the text and its format of exploration. In the last years, the evolution of the computer science profoundly modified our relationship to the text, leading to the development of new tools and analytical criteria. In this theoretical context, the concept of polyangular reading completes the notions of linear, reticular or matrix readings. This approach of the text became possible thanks to the always more powerful tools of corpus' edition proposed by text analysis environments, such as Hyperbase Web Edition. The polyangular reading principles rely both on the corpus based concepts and on the corpus driven concepts. It aims to multiply the partitions of the corpus to show as precisely as possible the thread of the textual fabric. This new method is then illustrated by two examples from the Latin bases developed by the LASLA of the University of Liege.

INDEX

Mots-clés : humanités numériques, Linguistique de corpus, Corpus based-corpus driven
Keywords : Digital Humanities, Corpus Linguistics, Corpus based-corpus driven

AUTEURS

MARGHERITA FANTOLI

LASLA – Université de Liège

MARC VANDERSMISSEN

LASLA – Université de Liège